Записки научных
семинаров ПОМИ
Том 552, 2026 г.

R. Oblovatny, A. Kuleshova, K. Polev, A. Zaytsev

PROBABILISTIC DISTANCES-BASED HALLUCINATION DETECTION IN LLMS WITH RAG

Abstract. Detecting hallucinations in large language models (LLMs) is critical for their safety in many applications. Without proper detection, these systems often provide harmful, unreliable answers. In recent years, LLMs have been actively used in retrieval-augmented generation (RAG) settings. However, hallucinations remain even in this setting, and while numerous hallucination detection methods have been proposed, most approaches are not specifically designed for RAG systems. To overcome this limitation, we introduce a hallucination detection method based on estimating the distances between the distributions of prompt token embeddings and language model response token embeddings[1]. The method examines the geometric structure of token hidden states to reliably extract a signal of factuality in text, while remaining friendly to long sequences. Extensive experiments demonstrate that our method achieves state-of-the-art or competitive performance. It also has transferability from solving the NLI task to the hallucination detection task, making it a fully unsupervised and efficient method with a competitive performance on the final task.

§1. Introduction

In recent years, large language models (LLMs) have been widely adopted in many applications. However, they often generate hallucinations — incorrect or fabricated content that does not match real-world facts or the provided context [1]. The latter case is of special interest, as it refers to incorrect generations in retrieval-augmented generation (RAG) settings, where LLMs rely on retrieved information to answer user queries. Since RAG is commonly used in chatbots [2] and other AI systems [3], detecting hallucinations in this setting is critical. Without proper detection, these systems may provide unreliable answers, which can be harmful in high-stakes fields such as law, finance, and healthcare.

Key words and phrases: hallucination detection, large language models, RAG systems, NLI.

While numerous hallucination detection methods have been proposed [4], most approaches are not specifically designed for RAG systems. Traditional methods often examine hallucinations in isolation, without considering how generated content relates to its supporting context. However, recent research [5] demonstrates that explicitly analysing the relationship between a model's generation and its provided context can significantly improve detection accuracy, as hallucinated answers consistently show weaker structural connections to their context in attention patterns compared to well-grounded responses.

We propose analysing this relationship in the space of LLM hidden states, as prior work has demonstrated that these states contain discriminative patterns of hallucinations [6, 7, 8]. Specifically, we hypothesize that the extent to which the response's embeddings deviate from those of the prompt may be indicative of hallucination, as they capture both text semantics and the model's latent "understanding" of the input [6]. To quantify this deviation, we employ probabilistic distances, such as Maximum Mean Discrepancy (MMD) [9].

Intuitively, we expect that embeddings of hallucinated examples should differ more from the context than grounded responses. Our experiments confirm this, demonstrating that distributions of hallucinated responses experience a shift in latent state space relative to the prompt. The observed tendency for probabilistic distances to increase for hallucination-inducing samples allows us to use these distances as an effective measure of hallucinations: the greater the distance, the higher the probability of hallucination. This simple but powerful approach demonstrates high detection efficiency, as evidenced by the results of our experiments (Tables 1, 2). In order for the score values to be comparable for sequences of different lengths and for the test to be valid in a non-asymptotic setting, we formulate the score using not the «raw» values of the statistics, but based on the *p-value* values for the corresponding null distributions.

The task of detecting hallucinations in the RAG setting is very similar to the NLI task, in which a contradiction between the hypothesis and the premise arises when the hypothesis contains new information that is absent from or opposite to the premise. Our method exploits this similarity and demonstrates good transferability between these tasks (Table 3), allowing it to be applied even when labeled examples of hallucinations are not available.

Our contributions are as follows:

(1) We analyze the probabilistic distances between the distributions of hidden states of prompts and responses in LLM, revealing that hallucinatory responses show greater deviations from prompts compared to grounded responses.
(2) Leveraging this observation, we propose a novel, model-intrinsic approach to hallucination detection that utilizes these distances between distributions to construct hallucination scores without requiring external knowledge bases or auxiliary models. Our procedure utilizes nonparametric Maximum Mean Discrepancy distances with wild bootstrap on top to get a principled hallucination score.
(3) Leveraging the similarity between the NLI task and the hallucination detection task, we construct an unsupervised hallucination detector. Its metrics are competitive, even with a significant domain shift from NLI to hallucination detection.
(4) Extensive experiments with five datasets and three LLMs up to 15B in size demonstrate that our method achieves state-of-the-art or competitive performance in hallucination detection.

§2. Related works

Hallucination detection. The problem of hallucinations in LLMs has attracted significant attention recently [10, 1, 11]. Existing methods can be roughly divided into several categories.

Consistency-based methods measure the homogeneity of multiple LLM answers to estimate uncertainty [12, 8, 13, 14, 15]. While these methods achieve reasonable detection performance, they require generating additional responses, making them computationally expensive. Moreover, their unsupervised nature inherently limits their effectiveness. *Uncertainty-based* methods rely on LLM token probabilities to assess model's confidence [16, 17]. Although efficient and easy to implement, they struggle to fully capture complex token dependencies [18] and are affected by biased token probabilities [19]. *Inner state-based* methods use hidden states or statistics from attention maps as input to lightweight classifiers [6, 7, 20].

A further limitation is that most methods overlook the RAG setting, despite it is very common in LLM applications [2]. Explicitly analyzing the relationship between the RAG context and the model's response could improve detection, as context-response relationships are known to indicate

hallucinations [5]. However, prior work focused on attention maps, leaving hidden state interactions underexplored.

Conclusion. Current hallucination detection methods exhibit several limitations when applied to retrieval-augmented generation (RAG) systems. The existing approaches fail to adequately model the complex statistical relationships between prompts and responses that could significantly improve detection accuracy. While previous work has primarily analyzed attention patterns in this regard [5], the rich semantic information encoded in transformer hidden states was overlooked. We address this gap by introducing a novel framework that employs probabilistic distances between prompt and response hidden state distributions to construct hallucination scores. We employ the NLI task to construct a fully-unsupervised version of the method that maintains the quality of the hallucination detection task.

§3. Preliminary

3.1. Attention mechanism. Modern LLMs are usually based on the self-attention mechanism [21]. This mechanism enables dynamic, context-aware token interactions by computing pairwise relevance scores across sequences. Each attention head operates as an independent feature detector, specializing in distinct linguistic or semantic patterns (e.g., syntactic roles or coreference) [22].

Formally, head h in layer ℓ generates a representation $\mathbf{e}_i^{(\ell,h)}$ for token i as follows:

$$\mathbf{e}_i^{(\ell,h)} = \sum_{j=1}^{n} \alpha_{ij}^{(\ell,h)} \mathbf{W}_V^{(\ell,h)} \mathbf{x}_j^{(\ell-1)},$$

where $\mathbf{x}_j^{(\ell-1)}$ is the hidden state of token j from layer $\ell - 1$ and

$$\alpha_{ij}^{(\ell,h)} = \operatorname{softmax}\left(\frac{(\mathbf{W}_Q^{(\ell,h)} \mathbf{x}_i^{(\ell-1)})^\top (\mathbf{W}_K^{(\ell,h)} \mathbf{x}_j^{(\ell-1)})}{\sqrt{d_k}} \right)$$

computes normalized attention weights, $\mathbf{W}_Q^{(\ell,h)}$, $\mathbf{W}_K^{(\ell,h)}$, $\mathbf{W}_V^{(\ell,h)}$ are learnable query, key, and value projection matrices, and d_k is the key dimension scaling factor.

3.2. Probabilistic distances. To quantify the relationship between retrieved inputs and generated responses, we propose analysing the statistical divergence between their respective hidden state distributions. This

approach is motivated by the hypothesis that hallucinated responses exhibit systematically different distributional characteristics (w.r.t. to the context) from grounded responses in the model's latent space. The choice of distance metric is crucial, as it must capture meaningful semantic differences while remaining robust to high-dimensional noise and accommodate the complex geometry of transformer hidden states. Probabilistic distances between distributions offer these properties while avoiding strong parametric assumptions about the underlying data. In our work, we employ the MMD distance measure.

Maximum mean discrepancy. The Maximum Mean Discrepancy (MMD) [9] is a statistical measure used to quantify the difference between two probability distributions based on samples drawn from them. It operates in a reproducing kernel Hilbert space (RKHS), where it computes the distance between the mean embeddings of the two distributions.

Given a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with characteristic kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the squared MMD between distributions α and β is defined as:

$$\begin{aligned}\mathrm{MMD}_k^2(\alpha, \beta) = \mathbb{E}_{\alpha\otimes\alpha}[k(x, x')] + \mathbb{E}_{\beta\otimes\beta}[k(y, y')] \\ - 2\mathbb{E}_{\alpha\otimes\beta}[k(x, y)].\end{aligned}$$

For two samples $X = \{x_i\}_{i=1}^m \sim \alpha,\ Y = \{y_j\}_{j=1}^n \sim \beta$, we consider the unbiased estimate of $\mathrm{MMD}_k(\alpha, \beta)$:

$$\begin{aligned}\widehat{\mathrm{MMD}}_u^2(X, Y) = \frac{1}{m(m-1)} \sum_{i\neq j}^{m} k(x_i, x_j) \\ + \frac{1}{n(n-1)} \sum_{i\neq j}^{n} k(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j),\end{aligned}$$

in our experiments.

Wild bootstrap. Under the null hypothesis $\alpha = \beta$, the asymptotic distribution of unbiased and biased estimates of MMD is degenerate, which complicates the analytical calibration of test thresholds on finite samples. Consequently, MMD is usually combined [23] with resampling-based procedures to obtain empirical p-values used as a score, suitable, e.g., for hallucination detection [24].

In this work, we use the wild bootstrap procedure to approximate the zero distribution of the MMD statistic and compute p-values. We use wild bootstrap instead of label shuffling because wild bootstrap allows us to

explicitly capture sequential dependence in the data. This is particularly important in our case, since the token-level latent representations we will use form structured sequences with strong local correlations caused by syntax, semantics, and the transformer architecture itself. Following [25, 26], for two sequences $X = \{x_i\}_{i=1}^m \sim \alpha$, $Y = \{y_j\}_{j=1}^n \sim \beta$, we generate a sequence of bootstrap weights $W := \{W_t\}_{t=1}^{n+m}$ according to a first-order autoregressive process,

$$W_t = \rho W_{t-1} + \sqrt{1-\rho^2}\,\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0,1), \tag{1}$$

where $\rho = \exp(-1/\ell)$ controls the temporal dependence of the weights, and ℓ is the bandwidth parameter. Given the weights W_t, the analogue of the MMD statistic for the wild bootstrap is obtained by weighting the kernel terms in the statistic,

$$\begin{aligned}\widehat{\mathrm{MMD}}^2_{w,u}(X,Y) = \frac{1}{m(m-1)}\sum_{i\neq j}^{m} \tilde{W}_i^{(x)}\tilde{W}_j^{(x)}k(x_i,x_j) + \\ + \frac{1}{n(n-1)}\sum_{i\neq j}^{n} \tilde{W}_i^{(y)}\tilde{W}_j^{(y)}k(y_i,y_j) \\ - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \tilde{W}_i^{(x)}\tilde{W}_j^{(y)}k(x_i,y_j),\end{aligned} \tag{2}$$

where $W_i^{(x)} = W_i$ for $i = \overline{1,m}$, $W_j^{(y)} = W_{j+m}$ for $j = \overline{1,n}$, $\tilde{W}_i^{(x)} = W_i^{(x)} - \frac{1}{m}\sum_k W_k^{(x)}$ and $\tilde{W}_j^{(y)} = W_j^{(y)} - \frac{1}{m}\sum_k W_k^{(y)}$ are centered versions of $W_i^{(x)}$ and $W_j^{(y)}$ respectively.

Given two samples $X = \{x_i\}_{i=1}^m$ and $Y = \{y_j\}_{j=1}^n$, for $b = 1, \ldots, B$ we generate a sequence of bootstrap weights $W^{(b)} := \{W_t^{(b)}\}_{t=1}^{n+m}$ and calculate $M^{(b)} := \widehat{\mathrm{MMD}}^2_{W^{(b)},u}(X,Y)$ using formula (2), obtaining an approximation of the null distribution $(W_b)_{b=1}^B$ of the test statistic. Through the bootstrap procedure, we obtain a *p-value*:

$$p\text{-}value := \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{1}\left[M^{(b)} \geq \widehat{\mathrm{MMD}}^2_u(X,Y)\right]\right), \tag{3}$$

which can be used to construct a normalized score.

§4. Methodology

4.1. Method. For a given prompt-response pair $[P, R]$, we propose using the MMD distance between the distributions of the hidden states of prompt tokens P and response tokens R as an indicator of hallucination: a larger distance $d(\alpha_P, \beta_R)$ indicates a higher probability of hallucination, where α_P and β_R denote the corresponding distributions of hidden states of tokens.

Since we want our test to be valid in the non-asymptotic framework and comparable for sequences of different lengths, we propose using the *p-value* based on the approximation of the null hypothesis using Monte Carlo approximation instead of the «raw» MMD value. Our experiments 7.2 demonstrate the advantage of this approach over the «raw» value of the statistic.

To achieve finer granularity, we extract hidden states from several model *heads* (selected by evaluating their discriminative power on the training dataset) rather than relying on standard layer-wise embeddings. Our experiments (Table 7) show that utilizing head embeddings improves detection quality, suggesting that faithfulness signals become "blurred" at the layer level.

Formally, our hallucination scores are formed as follows. Employing a set of pre-selected heads (see head selection procedure down below) H_{selected} of size N_{opt}, for a sample $S = [P, R]$ for each head $h \in H_{\text{selected}}$ we obtain the corresponding sequence of embeddings $[\mathbf{E}_P^h, \mathbf{E}_R^h] = [\mathbf{e}_{h,P}^1, \ldots, \mathbf{e}_{h,P}^{\text{len}(P)}, \mathbf{e}_{h,R}^1, \ldots, \mathbf{e}_{h,R}^{\text{len}(R)}]$. Given an estimate $\hat{d}$ of the probability distance d, which depends on the kernel function k, we use the bootstrap process to obtain the empirical null distribution of the test statistic d and the corresponding *p-value* using formula 3 above. The final score for head h is equal to $s_k(E_P^h, E_R^h) = 1 - p\text{-}value$. After calculating the scores for all heads, we average them:

$$\mathcal{HS}_k(S) = \frac{1}{N_{\text{opt}}} \sum_{h \in H_{\text{selected}}} s_k(\mathbf{E}_P^h, \mathbf{E}_R^h).$$

Our pipeline is briefly outlined in Algorithm 1. Below, we describe this procedure in detail.

Head selection. To select the attention heads, we adopt a procedure similar to the one proposed in [5]. Additionally, we compute hallucination scores using either MMD on the hidden states, rather than the topological divergence applied to attention maps.

The procedure assumes the availability of a annotated probe dataset V. For a fixed head h, we define the separation capability:

$$\Delta_h = \frac{1}{\#\{(P,R,y) \in V | y=1\}} \sum_{\{(P,R,y)\in V|y=1\}} s_k(E_P^h, E_R^h) - \frac{1}{\#\{(P,R,y) \in V | y=0\}} \sum_{(P,R,y)\in V|y=0\}} s_k(E_P^h, E_R^h), \tag{4}$$

where s_k we defined above. The procedure consists of three main steps:

(1) For each head h in the model, we measure the probabilistic distances across all training samples.
(2) We rank the heads by their separation capability Δ_h.
(3) From the top-performing heads, we select the optimal subset of size N_{opt} based on their joint performance (hallucination score as the average from the subset) measured on the validation set.

In line with [5], we constrain N_{opt} to a maximum of $N_{\max} = 6$ to maintain efficiency. An important advantage is that, as we show in experiments 6.2, the method demonstrates competitive performance when shifting the domain from NLI tasks to hallucination detection tasks, enabling its application in scenarios where labeled examples are unavailable.

§5. Experiments

5.1. Datasets. The proposed approach was evaluated on four datasets: RAGTruth [27], CoQA [28], SQuAD [29] and XSum [30]. The RAGTruth dataset contains manually annotated answers for several language models in RAG setting. The dataset includes annotations for samples for several tasks; in our paper, we experiment with two subsets of this dataset corresponding to the question answering (QA) and summarization (Summary) tasks.

As for the CoQA and SQuAD datasets, we employed the versions from the authors of the paper [5]. The authors of this paper used questions from the original datasets, generated answers using a language model, and annotated the resulting samples automatically using GPT-4o [31].

The statistics of the considered datasets are provided in Tables 8–9 (see Appendix A). Note that RAGTruth consists of samples with longer contexts and responses compared to SQuAD and CoQA, thus being more representative from the real-world performance point of view.

Algorithm 1 Head Selection

Require:
Score calculator s_k
Probe set $V = \{(P_i, R_i, y_j)\}$
Max heads N_{max}

Procedure HEADSELECTION
  for each head h_{ij} **do** ▷ Evaluate heads
    Calculate $\Delta_{h_{ij}}$ according to eq. 4
  end for
  $H \leftarrow \text{Sort}(\{h_{ij}\}, \text{key} = \Delta_{h_{ij}}, \text{descending})$
  $H_{\text{subset}} \leftarrow \emptyset$, $\text{AUROC}_{\max} \leftarrow 0$, $N_{\text{opt}} \leftarrow 1$
  for $N = 1$ **to** N_{max} **do**
    $H_{\text{subset}} \leftarrow H_{\text{subset}} \cup \{H_N\}$
    for each $(P_j, R_j, y_j) \in V$ **do**
      $p_j \leftarrow \frac{N-1}{N} p_j + \frac{1}{N} s_k(P_j^{H_N}, R_j^{H_N})$
    end for
    $\text{AUROC} \leftarrow \text{AUROC}(\{y_j\}, \{p_j\})$
    if $\text{AUROC} > \text{AUROC}_{\max}$ **then**
      $\text{AUROC}_{\max} \leftarrow \text{AUROC}$
      $N_{\text{opt}} \leftarrow N$
    end if
  end for
  return $H_{\text{subset}} = \{H_1, \ldots, H_{N_{\text{opt}}}\}$
end Procedure
$H_{\text{selected}} \leftarrow \text{HeadSelection}()$

5.2. Baselines. We compare our approach with nine established baselines spanning three methodological categories. For supervised hidden state-based methods, we evaluate SAPLMA, a linear probe over LM hidden states [6], and an attention-pooling probe that learns weighted combinations of layer activations [7]. The consistency-based methods include SelfCheckGPT measuring generation agreement via BERTScore [12], semantic entropy over clustered responses [13], and EigenScore analyzing variability in eigenspace [8]. The uncertainty-based methods comprise self-evaluation, perplexity, and maximum entropy [16].

5.3. Models. The method was evaluated on three popular open-source models: LLaMA-2-7B-chat [32], Mistral-7B-Instruct-v0.1 [33], and LLaMA-2-13B-chat [32].

5.4. Implementation details. For Maximum Mean Discrepancy, we used a Laplacian kernel with $l1$-norm as the kernel: $k(x, y) = \exp(-\frac{||x-y||_1}{\sigma})$, where σ is selected using median heuristics: $\sigma = \text{median}(||z_i - z_j||, i < j)$ for the pooled sequence $Z = [X, Y]$. The following parameters were set for wild bootstrap: bandwidth parameter l equals 1, bootstrap sample size equals 2048.

As for the baselines, we employed the following settings: for the consistency-based methods, we used $N = 20$ additional generations, following the ablation study in [12]. Hidden state-based classifiers were trained on the embeddings from the 16-th model layer, as the prior work has shown that the middle transformer layers contain the most factuality-related information [7, 6].

The reported results are averaged over 5 runs with different data splits, using test sets comprising 25% of the data and a fixed validation set size of 100.

§6. Results

6.1. Hallucination detection. The results of the experiments are presented in Tables 1–2. We compare the method with existing state-of-the-art methods for detecting hallucinations, both those requiring labeled samples and those that are fully unsupervised. The method demonstrates competitive results on datasets with long sequences (Table 1), achieving the best or second-best results, with the exception of the CNN/DM+Recent News dataset for the LLaMa-2-7B model.

The method demonstrates particularly strong performance on the largest of the presented models, Llama-2-13B. We also present results for the shorter sequence datasets CoQA and XSum in Table 2. The results degrade on these datasets, but still remain the best on Llama-2-13B, which may indicate that the method is more appropriate for larger models, exploiting the fact that large models better capture the factuality signal that MMD detects during hallucination detection.

6.2. Transfer from natural language inference task. The task of detecting hallucinations is very similar to the task of Natural Language Inference. We conduct experiments with head selection on the SNLI [34]

Method	Single generation	MS MARCO	CNN/DM + Recent News	SQuAD
Mistral-7B				
Semantic entropy	✗	**0.54** ± 0.03	0.51 ± 0.04	<u>0.70</u> ± 0.03
P(True)	✗	0.51 ± 0.04	**0.54** ± 0.03	0.44 ± 0.04
Semantic density	✗	<u>0.52</u> ± 0.04	<u>0.52</u> ± 0.03	0.50 ± 0.05
EigenScore	✗	**0.54** ± 0.04	0.50 ± 0.06	**0.71** ± 0.04
HaloScope	✓	0.57 ± 0.08	0.51 ± 0.10	<u>0.92</u> ± 0.07
LLM-Check	✓	0.49 ± 0.03	0.49 ± 0.03	0.50 ± 0.03
Perplexity	✓	0.45 ± 0.01	<u>0.54</u> ± 0.02	0.81 ± 0.05
Max entropy	✓	<u>0.68</u> ± 0.04	**0.60** ± 0.07	0.75 ± 0.05
ReDEEP	✓	0.54 ± 0.02	0.47 ± 0.06	0.45 ± 0.05
MMD(ours)	✓	**0.69** ± 0.03	**0.60** ± 0.05	**0.93** ± 0.01
LLaMA-2-7B				
Semantic entropy	✗	0.53 ± 0.03	0.51 ± 0.03	<u>0.73</u> ± 0.03
P(True)	✗	**0.57** ± 0.02	**0.54** ± 0.03	0.37 ± 0.03
Semantic density	✗	0.49 ± 0.03	0.45 ± 0.03	0.48 ± 0.03
EigenScore	✗	<u>0.55</u> ± 0.03	<u>0.53</u> ± 0.04	**0.75** ± 0.02
HaloScope	✓	0.51 ± 0.05	0.48 ± 0.05	0.67 ± 0.04
LLM-Check	✓	0.44 ± 0.02	0.49 ± 0.06	0.49 ± 0.01
Perplexity	✓	0.54 ± 0.04	0.44 ± 0.03	0.46 ± 0.03
Max entropy	✓	**0.65** ± 0.04	**0.59** ± 0.06	<u>0.73</u> ± 0.04
ReDEEP	✓	0.54 ± 0.04	0.52 ± 0.04	0.42 ± 0.08
MMD(ours)	✓	<u>0.64</u> ± 0.03	<u>0.53</u> ± 0.03	**0.78** ± 0.05
LLaMA-2-13B				
Semantic entropy	✗	**0.57** ± 0.04	**0.54** ± 0.03	<u>0.65</u> ± 0.03
P(True)	✗	0.55 ± 0.03	**0.54** ± 0.05	**0.84** ± 0.03
Semantic density	✗	0.52 ± 0.03	<u>0.52</u> ± 0.02	0.52 ± 0.02
EigenScore	✗	<u>0.56</u> ± 0.04	0.47 ± 0.04	0.57 ± 0.02
HaloScope	✓	0.54 ± 0.09	0.51 ± 0.04	0.55 ± 0.02
LLM-Check	✓	0.49 ± 0.06	**0.56** ± 0.05	0.57 ± 0.07
Perplexity	✓	0.54 ± 0.04	0.46 ± 0.07	0.45 ± 0.02
Max entropy	✓	<u>0.62</u> ± 0.03	0.53 ± 0.06	<u>0.78</u> ± 0.02
ReDEEP	✓	<u>0.62</u> ± 0.06	0.48 ± 0.05	0.48 ± 0.07
MMD(ours)	✓	**0.65** ± 0.03	<u>0.55</u> ± 0.07	**0.94** ± 0.02

Table 1. ROC AUC (↑) of hallucination detection techniques; best results in bold, second best underlined.

Method	Single generation	CoQA	XSum
Mistral-7B			
Semantic entropy	✗	**0.83** ± 0.02	0.70 ± 0.03
P(True)	✗	0.61 ± 0.03	0.44 ± 0.04
Semantic density	✗	0.56 ± 0.02	0.50 ± 0.05
EigenScore	✗	0.74 ± 0.02	**0.71** ± 0.04
HaloScope	✓	0.62 ± 0.08	0.92 ± 0.07
LLM-Check	✓	0.60 ± 0.01	0.50 ± 0.03
Perplexity	✓	0.54 ± 0.03	0.81 ± 0.05
Max entropy	✓	**0.73** ± 0.00	0.75 ± 0.05
ReDEEP	✓	0.59 ± 0.03	0.45 ± 0.05
MMD(ours)	✓	**0.73** ± 0.03	**0.93** ± 0.01
LLaMA-2-7B			
Semantic entropy	✗	**0.76** ± 0.01	0.73 ± 0.03
P(True)	✗	0.53 ± 0.03	0.37 ± 0.03
Semantic density	✗	0.52 ± 0.05	0.48 ± 0.03
EigenScore	✗	0.61 ± 0.03	**0.75** ± 0.02
HaloScope	✓	0.61 ± 0.04	0.67 ± 0.04
LLM-Check	✓	0.60 ± 0.03	0.49 ± 0.01
Perplexity	✓	**0.74** ± 0.02	0.46 ± 0.03
Max entropy	✓	0.65 ± 0.03	0.73 ± 0.04
ReDEEP	✓	0.72 ± 0.04	0.42 ± 0.08
MMD(ours)	✓	0.68 ± 0.04	**0.78** ± 0.05
LLaMA-2-13B			
Semantic entropy	✗	**0.76** ± 0.04	0.65 ± 0.03
P(True)	✗	0.67 ± 0.02	**0.84** ± 0.03
Semantic density	✗	0.48 ± 0.03	0.52 ± 0.02
EigenScore	✗	0.57 ± 0.03	0.57 ± 0.02
HaloScope	✓	0.57 ± 0.03	0.55 ± 0.02
LLM-Check	✓	0.57 ± 0.02	0.57 ± 0.07
Perplexity	✓	0.62 ± 0.03	0.45 ± 0.02
Max entropy	✓	0.66 ± 0.03	0.78 ± 0.02
ReDEEP	✓	0.73 ± 0.02	0.48 ± 0.07
MMD(ours)	✓	**0.84** ± 0.03	**0.94** ± 0.02

Table 2

dataset and transfer to datasets for hallucination detection. We expect the

method to behave similarly to natural language inference, since contradictions arise when there is a deviation from the given premise. Figure 1 shows the Spearman correlation between the resulting score and the SOTA NLI model roberta-large-mnli [35] on a test dataset for two models across LLM layers. The method demonstrates a strong stable correlation with the NLI score, which confirms the hypothesis that the method is sensitive to the emergence of new information in the hypothesis compared to the information presented in the premise. Table 3 presents the results of transferring our method with heads selected on the NLI task to datasets for hallucination detection. The method demonstrates competitive results even with a significant domain shift from the NLI task to hallucination detection. It is important to note that in this setting, the method does not require a labeled sample with hallucinations at all.

§7. Ablation study

7.1. What about another NLI datasets? We additionally conduct experiments on transfer from different NLI datasets to the hallucination detection task. The results are presented in Table 4. This table shows that the results can vary significantly across different datasets, suggesting that the selection of the NLI dataset is important for certain specific domains of the hallucination detection task. We assume that this depends on the structure of the NLI dataset and leave the question of its selection for a specific domain for future work.

7.2. Is bootstrap actually necessary? We compare our method with a similar approach, but in which we do not perform bootstrapping, but use the «raw» value of the MMD statistic to detect hallucinations. The results are presented in Table 5. As can be seen from the table, in most cases the bootstrap approach outperforms the «raw» statistic value, demonstrating more stable results. One of the advantages of using the bootstrap approach is that the resulting score is normalized from zero to one, while the «raw» statistic value can be within any range.

7.3. What about other approaches to bootstrapping? We conducted additional experiments with alternative approaches to bootstrapping under H_0: wild bootstrap with different methods of weight sampling, as well as label shuffling, which involves shuffling samples in the pooled sequence $Z = [X, Y]$ via permutation without replacement. The results are presented in Table 6. As can be seen from the table, wild bootstrap with

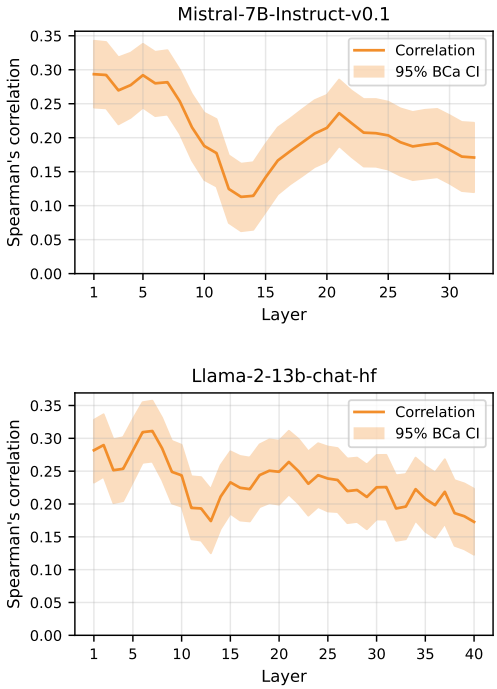

Figure 1. Spearman's rank correlation coefficient between $\mathcal{HS}_k$, calculated based on LLM layer embeddings, and the NLI score of the SOTA NLI model **roberta-large-mnli**. All experiments were conducted on a held-out test set. Top: Mistral-7B-Instruct-v0.1 embeddings. Bottom: Llama-2-13b-chat-hf embeddings.

AR weights consistently performs better than other methods, suggesting that this approach to weight sampling allows for accurate accounting of the autocorrelation of sample instances and is optimal when Maximum Mean Discrepancy is applied to samples of text token embeddings.

7.4. Do we need to consider head-level embeddings? To evaluate the necessity of fine-grained head selection, we compared our head-level MMD approach against a layer-level variant. For fair comparison, we

Method	MS MARCO	CNN/DM + Recent News	SQuAD
Mistral-7B			
Semantic entropy	**0.54** ± 0.03	0.51 ± 0.04	0.70 ± 0.03
P(True)	0.51 ± 0.04	**0.54** ± 0.03	0.44 ± 0.04
Semantic density	0.52 ± 0.04	0.52 ± 0.03	0.50 ± 0.05
EigenScore	**0.54** ± 0.04	0.50 ± 0.06	**0.71** ± 0.04
LLM-Check	0.49 ± 0.03	0.49 ± 0.03	0.50 ± 0.03
Perplexity	0.45 ± 0.01	0.54 ± 0.02	0.81 ± 0.05
Max entropy	0.68 ± 0.04	**0.60** ± 0.07	0.75 ± 0.05
MMD(snli)	**0.69** ± 0.03	0.59 ± 0.05	**0.86** ± 0.01
LLaMA-2-7B			
Semantic entropy	0.53 ± 0.03	0.51 ± 0.03	0.73 ± 0.03
P(True)	**0.57** ± 0.02	**0.54** ± 0.03	0.37 ± 0.03
Semantic density	0.49 ± 0.03	0.45 ± 0.03	0.48 ± 0.03
EigenScore	0.55 ± 0.03	0.53 ± 0.04	**0.75** ± 0.02
LLM-Check	0.44 ± 0.02	0.49 ± 0.06	0.49 ± 0.01
Perplexity	0.54 ± 0.04	0.44 ± 0.03	0.46 ± 0.03
Max entropy	**0.65** ± 0.04	**0.59** ± 0.06	**0.73** ± 0.04
MMD(snli)	**0.65** ± 0.04	0.55 ± 0.03	0.54 ± 0.03
LLaMA-2-13B			
Semantic entropy	**0.57** ± 0.04	**0.54** ± 0.03	0.65 ± 0.03
P(True)	0.55 ± 0.03	**0.54** ± 0.05	**0.84** ± 0.03
Semantic density	0.52 ± 0.03	0.52 ± 0.02	0.52 ± 0.02
EigenScore	0.56 ± 0.04	0.47 ± 0.04	0.57 ± 0.02
LLM-Check	0.49 ± 0.06	**0.56** ± 0.05	0.57 ± 0.07
Perplexity	0.54 ± 0.04	0.46 ± 0.07	0.45 ± 0.02
Max entropy	**0.62** ± 0.03	0.53 ± 0.06	0.78 ± 0.02
MMD(snli)	0.59 ± 0.06	0.53 ± 0.05	**0.94** ± 0.02

Table 3. ROC AUC (↑) of hallucination detection techniques during transfer MMD from the SNLI dataset to the hallucination detection task; best results in bold, second best underlined.

adapted the selection procedure from Algorithm 1 to identify optimal layers using the same training data. Results in Table 7 reveal that head-wise embeddings consistently outperform or performs on par with layer-wise ones, achieving higher ROC-AUC in the vast majority of cases.

NLI Dataset	MS MARCO	CNN/DM + Recent News	SQuAD	XSum
Mistral-7B				
SNLI [34]	**0.69** ± 0.03	0.59 ± 0.05	**0.86** ± 0.01	**0.54** ± 0.03
ANLI(round 1) [36]	0.68 ± 0.04	0.58 ± 0.06	0.85 ± 0.03	**0.54** ± 0.03
MNLI [37]	0.68 ± 0.02	**0.60** ± 0.05	0.41 ± 0.06	**0.54** ± 0.04
LLaMA-2-7B				
SNLI [34]	**0.65** ± 0.04	**0.55** ± 0.03	0.54 ± 0.03	**0.53** ± 0.03
ANLI(round 1) [36]	0.59 ± 0.03	**0.55** ± 0.04	**0.55** ± 0.04	0.47 ± 0.05
MNLI [37]	0.63 ± 0.03	**0.55** ± 0.03	0.40 ± 0.03	0.51 ± 0.06
LLaMA-2-13B				
SNLI [34]	0.59 ± 0.06	0.53 ± 0.05	**0.94** ± 0.02	**0.57** ± 0.03
ANLI(round 1) [36]	0.63 ± 0.04	**0.55** ± 0.05	0.92 ± 0.02	0.54 ± 0.05
MNLI [37]	**0.64** ± 0.03	0.53 ± 0.03	0.43 ± 0.03	0.55 ± 0.05

Table 4. ROC AUC of MMD for three LLMs during transfer from head selection on the NLI dataset for different datasets. The best results for each model are highlighted in bold.

The advantage is most obvious on the SQuAD dataset, where head-level MMD exceeds layer-level performance by over 0.2 ROC-AUC for Llama-2-7B and 0.1 for Mistral-7B. This demonstrates that head-specific embeddings better preserve factuality signals, while layer aggregation appears to dilute discriminative information.

7.5. How large should the probe sets be? To evaluate the sensitivity of our method to probe set size, we conducted a study of the dependence of the results on this factor. The results presented in Table 2a) confirm that the method is practically independent of probe set size and maintains quality even with smaller sizes.

7.6. How many heads are needed? We conducted a study of the dependence of the method's performance on the number of heads $N_{\max}$. The results are shown in Figure 2b). It can be seen that the method does not depend significantly on the number of heads, and $N_{\max} = 6$ is the

Version	MS MARCO	CNN/DM + Recent News	CoQA	SQuAD	XSum
Mistral-7B					
Bootstr.	**0.69** $\pm$ 0.03	**0.60** $\pm$ 0.05	**0.73** $\pm$ 0.03	**0.93** $\pm$ 0.01	0.59 $\pm$ 0.04
Raw	0.67 $\pm$ 0.02	0.53 $\pm$ 0.06	0.70 $\pm$ 0.03	0.88 $\pm$ 0.01	**0.61** $\pm$ 0.03
LLaMA-2-7B					
Bootstr.	**0.64** $\pm$ 0.03	**0.53** $\pm$ 0.03	0.68 $\pm$ 0.04	**0.78** $\pm$ 0.05	0.55 $\pm$ 0.06
Raw	0.55 $\pm$ 0.03	0.46 $\pm$ 0.03	**0.74** $\pm$ 0.02	0.76 $\pm$ 0.05	**0.56** $\pm$ 0.02
LLaMA-2-13B					
Bootstr.	**0.65** $\pm$ 0.03	**0.55** $\pm$ 0.07	0.84 $\pm$ 0.03	**0.94** $\pm$ 0.02	**0.60** $\pm$ 0.02
Raw	0.56 $\pm$ 0.04	0.48 $\pm$ 0.03	**0.85** $\pm$ 0.04	0.92 $\pm$ 0.02	0.59 $\pm$ 0.07

Table 5. ROC-AUC scores for hallucination detection comparing versions with and without bootstrapping.

optimal number. A larger number of heads can lead to signal noise, while a smaller number can lead to insufficient robustness.

§8. Conclusion

We propose a novel method for detecting hallucinations based on maximum mean divergence (MMD), which measures the divergence between the distributions of the query and response hidden states. To improve detection quality, we use attention-head embeddings rather than layer embeddings, as our ablation studies confirm that layer-level representations blur factuality signals. In addition, we use bootstrapping approaches to account for the autoregressiveness of text tokens. We further demonstrate that the method does not require a large amount of labeled data in its standard setting, nor does it require a large number of heads to capture the factuality signal.

Notably, the method has a strong correlation with the solution to the NLI task and demonstrates good transferability with a strong shift to the hallucination detection domain, allowing the method to operate in a fully unsupervised mode.

Bootstrap type	MS MARCO	CNN/DM + Recent News	SQuAD
Mistral-7B			
Label shuffling	0.67 ± 0.02	$\underline{0.59} \pm 0.04$	$\underline{0.92} \pm 0.02$
Wild with iid normal weights	$\underline{0.68} \pm 0.02$	$\underline{0.59} \pm 0.04$	$\mathbf{0.93} \pm 0.01$
Wild with iid rademacher weights	0.67 ± 0.03	$\underline{0.59} \pm 0.04$	$\mathbf{0.93} \pm 0.02$
Wild with AR(1) weights (ours)	$\mathbf{0.69} \pm 0.03$	$\mathbf{0.60} \pm 0.05$	$\mathbf{0.93} \pm 0.01$
LLaMA-2-7B			
Label shuffling	$\underline{0.62} \pm 0.02$	$\mathbf{0.56} \pm 0.03$	0.65 ± 0.05
Wild with iid normal weights	$\mathbf{0.64} \pm 0.03$	$\underline{0.54} \pm 0.03$	0.72 ± 0.03
Wild with iid rademacher weights	$\mathbf{0.64} \pm 0.03$	$\underline{0.54} \pm 0.05$	$\underline{0.75} \pm 0.07$
Wild with AR(1) weights (ours)	$\mathbf{0.64} \pm 0.03$	0.53 ± 0.03	$\mathbf{0.78} \pm 0.05$
LLaMA-2-13B			
Label shuffling	$\underline{0.63} \pm 0.04$	$\underline{0.52} \pm 0.04$	$\underline{0.94} \pm 0.02$
Wild with iid normal weights	$\mathbf{0.65} \pm 0.04$	$\underline{0.52} \pm 0.05$	$\mathbf{0.95} \pm 0.02$
Wild with iid rademacher weights	$\mathbf{0.65} \pm 0.04$	0.51 ± 0.05	$\underline{0.94} \pm 0.02$
Wild with AR(1) weights (ours)	$\mathbf{0.65} \pm 0.03$	$\mathbf{0.55} \pm 0.07$	$\underline{0.94} \pm 0.02$

Table 6. ROC AUC (↑) of MMD-based hallucination detection techniques for three LLMs for different bootstrap types. The best results for each model are highlighted in bold, and the second best are underlined.

Our method achieves state-of-the-art or competitive results on various tests, especially on larger models. Our results establish a simple but effective geometric approach to hallucination detection based on statistical divergence of latent representations.

Appendix A. Datasets

Tables 8 and 9 provide comprehensive statistics for the considered datasets. Table 8 presents the distribution of hallucinated versus grounded examples, while Table 9 details the average prompt and response lengths across datasets.

The data reveals distinct characteristics among the datasets: CoQA and SQuAD contain samples with relatively short model responses, whereas RAGTruth QA and RAGTruth Summ present more challenging cases with

Granularity	MS MARCO	CNN/DM + Recent News	SQuAD	XSum
Mistral-7B				
Heads	**0.69** ± 0.03	**0.60** ± 0.05	**0.93** ± 0.01	**0.59** ± 0.04
Layers	0.67 ± 0.02	**0.60** ± 0.03	0.83 ± 0.01	0.55 ± 0.03
LLaMA-2-7B				
Heads	**0.64** ± 0.03	**0.53** ± 0.03	**0.78** ± 0.05	0.55 ± 0.06
Layers	0.63 ± 0.04	0.52 ± 0.03	0.55 ± 0.04	**0.56** ± 0.05
LLaMA-2-13B				
Heads	**0.65** ± 0.03	**0.55** ± 0.07	**0.94** ± 0.02	**0.60** ± 0.02
Layers	0.58 ± 0.03	0.52 ± 0.06	0.92 ± 0.02	0.51 ± 0.03

Table 7. ROC-AUC scores for hallucination detection comparing head-wise and layer-wise embeddings.

Model	**RAGTruth QA**		**RAGTruth Summ**		**SQuAD**		**CoQA**	
#	Hal.	Gr.	Hal.	Gr.	Hal.	Gr.	Hal.	Gr.
Mistral-7B	378	561	615	325	311	389	776	776
LLaMA-2-7B	497	474	434	509	357	235	375	375
LLaMA-2-13B	396	588	276	623	314	436	279	384

Table 8. Dataset statistics showing the number of hallucinated (Hal.) and non-hallucinated (Gr., grounded) examples for each model.

both lengthy prompts and detailed responses. This increased complexity makes RAGTruth particularly valuable for evaluating method performance under conditions that closely resemble real-world applications.

All methods were tested on isolated subsets of the datasets, which were split into training and test subsets following the authors of the original RAGTruth paper [27] and the annotated versions of the SQuAD and CoQA datasets from [5].

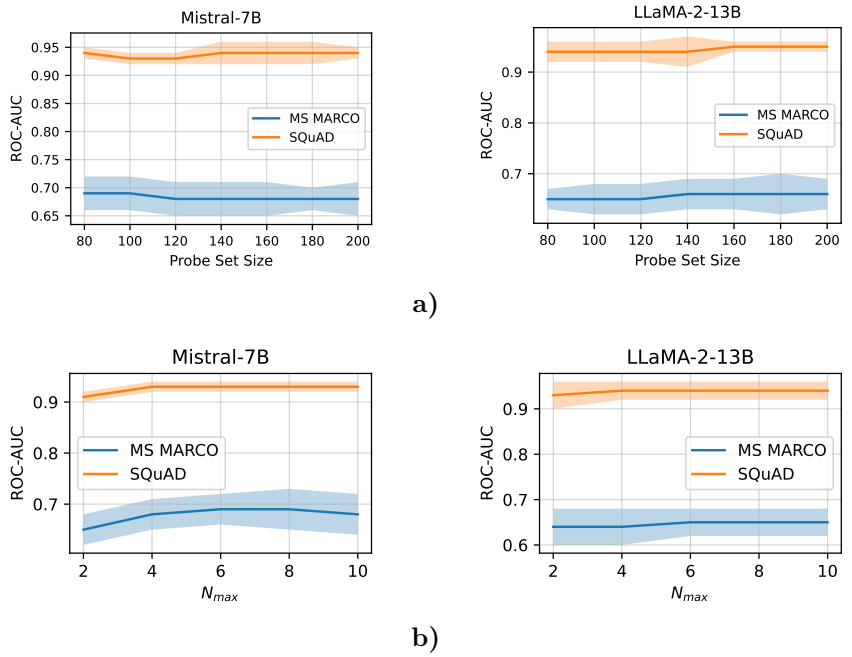

Figure 2. ROC-AUC performance on Mistral-7B(left) and Llama-2-13B(right) across **a)** different numbers of selected attention heads ($N_{\max}$) and **b)** different probe set sizes with fixed $N_{\max} = 6$.

Model	**RAGTruth QA**		**RAGTruth Summ**		**SQuAD**		**CoQA**	
Length	Pr.	Resp.	Pr.	Resp.	Pr.	Resp.	Pr.	Resp.
Mistral-7B	431	143	814	156	256	17	523	9
LLaMA-2-7B	446	267	848	152	258	34	546	10
LLaMA-2-13B	447	213	784	142	292	17	548	10

Table 9. Dataset statistics. Average prompt length (Pr.) and response length (Resp.) for all models.

Appendix B. Computational resources

All experiments were run on a machine equipped with one NVIDIA A100 GPU (40 GB) and one NVIDIA L100 GPU (96 GB).

References

1. L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, *et al.*, *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*, ACM Transactions on Information Systems (2023).
2. R. Akkiraju, A. Xu, D. Bora, T. Yu, L. An, V. Seth, A. Shukla, P. Gundecha, H. Mehta, A. Jha, *et al.*, *FACTS About Building Retrieval-Augmented Generation-Based Chatbots*, ArXiv preprint arXiv:2407.07858 (2024).
3. M. Kothapalli, *The Practical Applications of Retrieval-Augmented Generation in AI*, JCSSD **3** (2024).
4. P. Sahoo, P. Meharia, A. Ghosh, S. Saha, V. Jain, and A. Chadha, *A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models*, in: Findings of the Association for Computational Linguistics: EMNLP, 2024, pp. 11709–11724.
5. A. Bazarova, A. Yugay, A. Shulga, A. Ermilova, A. Volodichev, K. Polev, J. Belikova, R. Parchiev, D. Simakov, M. Savchenko, *et al.*, *Hallucination Detection in LLMs via Topological Divergence on Attention Graphs*, ArXiv preprint arXiv:2504.10063 (2025).
6. A. Azaria and T. Mitchell, *The Internal State of an LLM Knows When It's Lying*, in: Findings of the Association for Computational Linguistics: EMNLP, 2023, pp. 967–976.
7. C.-W. Sky, B. Van Durme, J. Eisner, and C. Kedzie, *Do Androids Know They're Only Dreaming of Electric Sheep?*, in: Findings of the Association for Computational Linguistics: ACL, 2024, pp. 4401–4420.
8. C. Chen, K. Liu, Z. Chen, Y. Gu, Y. Wu, M. Tao, Z. Fu, and J. Ye, *INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection*, in: The Twelfth International Conference on Learning Representations, 2024.
9. A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, *A Kernel Two-Sample Test*, Journal of Machine Learning Research **13** (2012), 723–773.
10. Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, *et al.*, *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models*, ArXiv preprint arXiv:2309.01219 (2023).
11. Y. Wang, M. Wang, M. A. Manzoor, F. Liu, G. Georgiev, R. Das, and P. Nakov, *Factuality of Large Language Models: A Survey*, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 19519–19529.
12. P. Manakul, A. Liusie, and M. Gales, *SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models*, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023.
13. L. Kuhn, Y. Gal, and S. Farquhar, *Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation*, in: The Eleventh International Conference on Learning Representations, 2023.
14. X. Qiu and R. Miikkulainen, *Semantic Density: Uncertainty Quantification for Large Language Models through Confidence Measurement in Semantic Space*, in: Advances in Neural Information Processing Systems, 2024.

15. A. Nikitin, J. Kossen, Y. Gal, and P. Marttinen, *Kernel Language Entropy: Fine-Grained Uncertainty Quantification for LLMs from Semantic Similarities*, Advances in Neural Information Processing Systems **37** (2024), 8901–8929.
16. E. Fadeeva, A. Rubashevskii, A. Shelmanov, S. Petrakov, H. Li, H. Mubarak, E. Tsymbalov, G. Kuzmin, A. Panchenko, T. Baldwin, *et al.*, *Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification*, ArXiv preprint arXiv:2403.04696 (2024).
17. A. Malinin and M. Gales, *Uncertainty Estimation in Autoregressive Structured Prediction*, in: International Conference on Learning Representations, 2021.
18. D. N. Yaldiz, Y. F. Bakman, B. Buyukates, C. Tao, A. Ramakrishna, D. Dimitriadis, J. Zhao, and S. Avestimehr, *Do Not Design, Learn: A Trainable Scoring Function for Uncertainty Estimation in Generative LLMs*, in: Findings of the Association for Computational Linguistics: NAACL, 2025.
19. I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed, *Bias and Fairness in Large Language Models: A Survey*, Computational Linguistics (2024).
20. Y.-S. Chuang, L. Qiu, C.-Y. Hsieh, R. Krishna, Y. Kim, and J. Glass, *Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps*, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 1419–1436.
21. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, *Attention Is All You Need*, in: Advances in Neural Information Processing Systems, 2017.
22. E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, *Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned*, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5797–5808.
23. A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton, *MMD Aggregated Two-Sample Test*, Journal of Machine Learning Research **24** (2023), 1–81.
24. S. Rabanser, S. Günnemann, and Z. Lipton, *Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift*, Advances in Neural Information Processing Systems **32** (2019).
25. X. Shao, *The Dependent Wild Bootstrap*, Journal of the American Statistical Association **105** (2010), 218–235.
26. K. Chwialkowski, D. Sejdinovic, and A. Gretton, *A Wild Bootstrap for Degenerate Kernel Tests*, Advances in Neural Information Processing Systems **27** (2014).
27. C. Niu, Y. Wu, J. Zhu, S. Xu, K. Shum, R. Zhong, J. Song, and T. Zhang, *RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models*, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 10862–10878.
28. S. Reddy, D. Chen, and C. D. Manning, *CoQA: A Conversational Question Answering Challenge*, Transactions of the Association for Computational Linguistics **7** (2019), 249–266.

29. P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, *SQuAD: 100,000+ Questions for Machine Comprehension of Text*, in: Conference on Empirical Methods in Natural Language Processing, 2016.
30. S. Narayan, S. B. Cohen, and M. Lapata, *Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization*, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 1797–1807.
31. A. Hurst, A. Lerer, A. P. Goucher, *et al.*, *GPT-4o System Card*, ArXiv preprint arXiv:2410.21276 (2024).
32. H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, *Llama 2: Open Foundation and Fine-Tuned Chat Models*, ArXiv preprint arXiv:2307.09288 (2023).
33. D. S. Chaplot, A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. de las Casas, F. Bressand, G. Lengyel, *et al.*, *Mistral-7B*, ArXiv preprint arXiv:2310.06825 (2023).
34. S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, *A Large Annotated Corpus for Learning Natural Language Inference*, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 632–642.
35. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, ArXiv preprint arXiv:1907.11692 (2019).
36. Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, *Adversarial NLI: A New Benchmark for Natural Language Understanding*, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
37. A. Williams, N. Nangia, and S. Bowman, *A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference*, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 1112–1122.

Поступило 4 февраля 2026 г.

Markov Lab,
Department of Mathematics and Computer Science,
St. Petersburg University, St. Petersburg, Russia
E-mail: oblovatnyi@gmail.com

AI Center, Skoltech, Moscow, Russia
E-mail: A.Bazarova@skoltech.ru

SB AI Lab, Moscow, Russia
E-mail: endless.dipole@gmail.com

AI Center, Skoltech, Risk Department,
Sber, Moscow, Russia
E-mail: A.Zaytsev@skoltech.ru